Bridging the Gap between Plato and His Successors: Towards an Annotated Gold Standard of Intertextual References to Plato in Ancient Greek Literature

Marcus Pöckelmann  
marcus.poeckelmann@informatik.uni-halle.de  
Institute of Computer Science  
Martin Luther University Halle-Wittenberg

Eva Wöckener-Gade  
woeckener-gade@uni-leipzig.de  
Institut für Klassische Philologie und Komparatistik  
Leipzig University

Introduction

The reception of Plato’s work in ancient literature is characterized by a tremendous diversity and has not nearly been researched sufficiently. Our research project aims to further reveal his impact on Ancient Greek literature, particularly by finding paraphrases in an interdisciplinary approach combining expertise from humanities, computational linguistics, and computer science. The research project is ambitious in terms of methodology: it explores which algorithms for paraphrase identification are suited for smaller volumes of text beyond Big Data, and how to adjust their parameters to yield valuable results for literary phenomena which are less frequent and partly more elaborate than those in everyday language. A crucial precondition to reach this goal is a thorough analysis of different types of paraphrases in Ancient Greek literature that grants us an understanding of what we are searching for. On the other hand, we need a stable basis to evaluate our approaches to paraphrase identification by validating our results and to ensure comparability of different algorithms and their combinations. For both demands we could not resort to parallel corpora and similar resources that already exist for many modern languages. In our presentation we want to introduce our gold standard of over 200 intertextual references to Plato as well as the annotation tool we developed to characterize and categorize these references. We consider the annotated gold standard as a starting point for a new view on the concept of paraphrasing adjusted to the actualities of ancient literature as well as a touchstone for the development and the systematic evaluation of our algorithms for finding yet undiscovered paraphrases.

Selection of references: our gold standard

To compile a representative selection of paraphrases of Plato, we had to meet some challenging demands: first to cover a timespan of almost 1000 years of Ancient Greek literature with its linguistic changes, and second to include works from different genres and authors with divergent approaches to Plato. Moreover, it was required to comprise different types of intertextual references to meet with our wider definition of 'paraphrase', ranging from almost verbatim representations to instances characteristic of literature as e.g. allusions with a paraphrastic core that show only minor verbal resemblances to the platonic passage. We tried to bridge these difficulties by making a diachronic cross section through our corpus, the Thesaurus Linguae Graecae (TLG): we selected several crucial authors from different periods like Aristotle (4th century BC), Plutarch (1st-2nd century AD), up to Olympiodorus (6th century AD) but included also minor ones. For the identification of the references we could draw upon the preliminary work of different classical scholars, using major text editions with *apparatus fontium* and an index of intertextual references to Plato’s *Republic* (Boter 1989). The final document includes 216 passages of Plato with the corresponding paraphrase in Greek, each coupled with CTS-URNs and German translation.

Canonical Text Services

We use CTS-URNs – Uniform Resource Names defined by the Canonical Text Services (developed by Blackwell and Smith (2014), considering some special demands of ancient texts) – for referencing passages in our corpus, with the following adaptations for our project: To pin down paraphrases, an exact referencing is required not only on a scale of text segments or lines but of words. Therefor all words of a line are numbered consecutively and thus the position of the word rather than the word itself functions as the point of reference.

2 Cf. Bhagat and Hovy (2013) and Vila et al. (2014) for the need and difficulties of a broader view on paraphrases.
3 We are using the TLG-E version, reformatted to XML, cleaned and enhanced with annotations.
Annotating intertextual references: our ‘Referenzannotierer’

To analyze and visualize the paraphrases of our gold standard and newfound instances of intertextual references, we developed an interactive web-based tool, the so called Referenzannotierer (RA). The RA allows a general categorization of the paraphrase plus the annotation of its word-to-word-relations. It is conceived as a flexible yet consistent low-threshold system based on comparison of syntactical and lexical rather than semantic criteria to avoid biased judgements as far as possible. Short explanations are embedded for the different segments of the RA, easily accessible by mouseover. Although conceived for the annotation of paraphrases in Ancient Greek, the RA is apt for other forms of intertextual references, e.g. citations and allusion as well as other languages. Different students and fellow-researchers have successfully tested it.

The RA is organized in three parts:

In the upper part the given reference can be characterized in different categories that are based on a synthesis of theories from linguistic and literary studies and our own theoretical approach to the concept of paraphrase in Greek literature: The relation of the passages is described regarding additions/subtractions (cf. Plett 1991, 3-29). Moreover, the degree of similarity on the lexical level and the marking of the reference can be categorized as non-existent, implicit, or explicit (cf. Helbig 1996, 83-135).

In the middle part of RA the two text passages of the reference can be selected via CTS-URNs. The passage from Plato (pretext) on the left and the corresponding passage from a later author (posttext) on the right are displayed as a continuous text with some context. They are also linked to a full text representation.

The core of the RA is its lower part where different types of relations between single words can be annotated. For that purpose, the two text passages are displayed a second
time as strings opposite each other and the word-to-word-relations are visualized as lines in the free space between them. Relations illustrating similarities and differences between the passages such as “same lemma, same form (=citation)” as well as additions to/subtractions from the pretext can be selected from a pop-up dialog. The visualization of these is enhanced by different features like graduation of colors and shades and highlighting on mouseover. Numbers retrieved by similarity measurement based on a word2vec\(^4\)-embedding of the corpus further classify the semantic proximity of the single relations between pre- and posttext.

Figure 4: the lower part of RA: annotation of word-to-word relations between pre- and posttext. Note that the annotations are highlighted in the middle part too.

**Annotating the gold standard with the RA**

With our annotated gold standard, we compiled a comprehensive and representative collection of analyzed and categorized intertextual references to Plato within the Ancient Greek literature. All annotators had excellent command of Ancient Greek language as demanded by the task. We established inter-annotator agreement for various problems; the resulting guidelines made their way into the annotation tool as embedded explanations. The annotations were revised in two steps (once by the original annotator and once more by an independent one). We will present statistics of datasets retrieved from the annotations that grant new insights into the concept of paraphrasing especially but not only in Ancient Greek literature. For instance, we can correlate the percentages of verbatim reproductions in the posttext with our different categories of paraphrases. We suggest, that such results can be put to use as supplements for existing approaches to semi-automatic paraphrase identification like those pursued in our project.

\(^4\)Mikolov et al. (2013).
Access and licenses

Most of the project's tools and data are still within an active process of development. Therefore, a release (under a free license) is scheduled near the end of the project. An online demonstrator of the annotator will be made available for the conference.

References


